Predicting Substrates of $\gamma-$secretase in *Drosophila*

by

Yuhao Wang

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author .....
Department of Electrical Engineering and Computer Science
May 20, 2016

Certified by .....
Bonnie Berger
Professor of Mathematics and EECS
Thesis Supervisor

Accepted by .....
Professor Leslie A. Kolodziejski
Chair, Department Committee on Graduate Theses

Predicting Substrates of $\gamma-$secretase in *Drosophila*

by

Yuhao Wang

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2016, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

In this thesis, I designed a computational method for predicting substrates of $\gamma-$secretase in *Drosophila* and compared this method with other popular methods using some benchmark data set. Results show our method significantly outperforms other methods and generates important candidates for further experimental validation. Results from this computational experiment could be very important in comparing two common hypotheses for Alzheimer's disease: the presenilin hypothesis and amyloid hypothesis.

Thesis Supervisor: Bonnie Berger
Title: Professor of Mathematics and EECS

Acknowledgments

It has been my great honour to work with brilliant faculty and students in MIT. Several years ago, when I started doing research as undergraduate research assistant in Tsinghua University, I would never imagine I could have a chance to work in such top research institute for my research career. I sincerely appreciate the great support from my thesis supervisor, Professor Bonnie Berger, and also Professor Jian Peng at UIUC on this project, they have provided many insightful suggestions and helped me to move forward when I was in trouble. I also feel deeply thankful for our biological collaborators, Professors Jie Shen and Norbert Perrimon at Harvard for their help on the biological background of this project. I thank Dr. Sepehr Ehsani in our lab for his patience in explaining sophisticated biological concepts in protein folding, as well as his philosophical insight in biology. I am also deeply thankful for the support from the other lab members in Berger lab for their constructive suggestions in this project, as well as mental support from my friends and family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Inherited mutations in genes encoding presenilins (PS) and amyloid precursor proteins (APP) are strongly associated with Alzheimer's disease, but the molecular mechanism is still unknown. As presenilins could form γ-secretase, a protease complex involved in the cleavage of many type I transmembrane proteins [1], such as Notch and N-cadherin, and APP is a substrate of $\gamma-$secretase, two hypothese for the molecular mechanism exist.

The first is the amyloid hypothesis [2, 3], which states that overexpression of APP triggers neurodegeneration and thus leads to Alzheimer's disease (AD); here, Alzheimer's associated mutations in PS are mainly caused by cleavage of APP from $\gamma-$secretase. The second is the presenilin hypothesis [4], which states that degeneration of presenilin causes accumulation of its substrates, mainly the type I transmembrane proteins, and eventually leads to neurodegeneration; here, APP may be a factor but not the main cause. There is also experimental evidence that supports this latter hypothesis in mouse [5] with presenilin knockout in the forebrain.

In this project we are primarily interested in investigating which hypothesis is correct. In order to verify the two hypotheses, the first thing we need to do is to find all the substrates of $\gamma-$secretase for experiments to test association of each substrate with AD, which is also the primary focus of this project. It is a very important step to

test the two hypotheses as after that, we could be able to compare the two hypotheses by testing association of each substrate predicted from the first step with AD in real experiment, which would be further discussed in Sec. 1.2.2.

As experiments to find the target of $\gamma-$secretase in mice or other higher organisms are not practically feasible due to time and cost, using *Drosophila* as a model system has the benefit of its short lifespan, feasibility of phenotypic screening, and huge resources (eg. Genome-wide transgenic RNAi library [6]). Once we identify candidate genes in the fly system, we will validate them in the mouse or human systems with our experimental collaborators Dr. Jie Shen and Norbert Perrimon.

Traditionally researchers would select a list of proteins from empirical experience and experimentally detect their association with Alzheimer's disease; as the number of protein candidates is too large [1], it is important to develop data analysis techniques to prioritize proteins for experimental testing.

The entire project is composed of two parts: the first is to use computational approaches to discover transmembrane proteins cleaved by $\gamma-$secretase, and the second, experimental approaches to identify its association with Alzheimer's disease.

1.2 Experimental Methods for Alzheimer's Associated Protein Discovery

The experiments are based on RNA interference (RNAi) [7], a technology that uses RNA molecules to inhibit gene expression [8].

1.2.1 RNA Interference for Gene Knockdowns

RNA interference technology is based on the phenomenon that some RNA molecules can destruct specific mRNA to suppress mRNA expressions. It was first observed in Fire et al. [9] where *Caenorhabditis elegans* used interfering RNA to cleave double-stranded RNA for genetic regulation.

Two kinds of RNA are used for interfering with RNA: microRNAs (miRNA) and

small interfering RNAs (siRNA); each recognizes specific mRNA by their sequence and silence specific RNA based on their sequences.

Based on this property, RNAi is widely used as a gene silencing technique; this technique is achieved by the following procedure:

- Introduction of siRNA into the cell, either through direct introduction or using an expression vector [10] to synthesize siRNA within the cell;
- RNA-induced silencing complex (RISC) [11] processes the siRNA and degrades mRNAs complementary to the induced siRNA.

1.2.2 Experiment to Compare Presenilin Hypothesis and Amyloid Hypothesis

Here is how we design an experiment to compare the presenilin and amyloid hypotheses.

For all the flies we are going to use in this experiment, we use RNAi to knockdown presenilin; then for each substrate of $\gamma-$secretase, we have a perspective group of fly where this substrate is also knocked down using RNAi, so that we can distinguish the association between the substrate and Alzheimer's disease by observation of double knockdowns, if AD symptoms are weakened or strengthened as a result of double knockdowns, then we say this substrate is associated with AD; if double knockdowns does not cause any changes relative to single knockdown, then the substrate is not associated with AD.

There are many ways to detect the association between the substrate and AD, one is through lifespan; lifespans for flies with more serious AD symptoms would be shorter, and vice versa. As measurements using lifespan are quite unstable, we also consider using microscopy technologies.

We pursue the intuition that if there are a lot of substrates verified to be associated with AD, then the presenilin hypothesis is correct, otherwise, the amyloid hypothesis is correct. As one can imagine finding the substrates of $\gamma-$secretase is very important

to initialize the experiment. As the substrate of $\gamma-$secretase is still unknown, we need to design computational methods to discover them.

1.3 Computational Framework for Protein Interaction Prediction

As stated in Sec. 1.2, from a computational perspective, to test the hypothesis, we need to design computational methods to predict substrates of $\gamma-$secretase. the motif based method has been widely used to predict interaction candidates of proteins. The core part, motif finding, can be split into two approaches, one using multiple sequence alignment and constructing motif using alignment results (alignment based method); another non-alignment based method, where we construct a motif without sequence alignment.

1.3.1 Alignment Based Method for Motif Finding

The alignment based method has been widely used in constructing motifs for a wide variety of biological sequences; for example, Lambert et al. [12] used RNA multiple sequence alignment with secondary structure information to find RNA motifs and Kellis et al. [13, 14] used whole genome alignment of multiple yeast species to find regulatory motif elements across genome.

Although alignment based methods have been successfully applied to motif finding of RNA and DNA sequences, it has been a challenge to apply these to finding protein sequence motifs. Challenges involve:

- protein function is mostly determined by structure, so discovering sequence motifs is not so helpful for understanding function;
- alphabet size of amino acids is much larger than nucleotides, which significantly reduces multiple sequence alignment performance.

Here we overcome these challenges. For the first challenge, as $\gamma-$secretase would cleave at a transmembrane region, where proteins are unfolded into $\alpha-$helix structure, which is very close to their native state: linear sequences from a computational perspective, sequence based methods would be more informative than structure based methods. For the second, we could use a recently proposed alphabet reduction method [15] to improve alignment quality.

1.3.2 Non-alignment Based Method for Motif Finding

There are also other motif finding approaches without alignment, the most famous one would be EM based motif finding algorithm [16], like MEME [17] and DREME [18]. They both assume a generative model that generates the sequences for protein interaction and use EM-based algorithm for parameter estimation.

1.4 Thesis Overview

The rest of thesis is composed of the following parts. Chapter 2 contains the computational method for predicting $\gamma-$secretase substrates, Chapter 3 contains the results analysis and comparison with other popular methods, and Chapter 4, a discussion chapter as well as topics for future research.

Chapter 2

Methods

In this chapter we give a detailed summary of the method designed to discover substrates of $\gamma-$secretase. The whole method is based on motifs. First, we used multiple bioinformatics tools to extract homologous proteins of the known substrates of $\gamma-$secretase, then use these proteins, together with the known substrates of $\gamma-$secretase, as input data to train the motif model. After we have trained a motif, we use the motif scanning algorithm to find the substrates of $\gamma-$secretase.

In this chapter, we will start by introducing the bioinformatics methods we are going to use to train the motif, including homologous protein discovery for training data for the motif model (Sec. 2.1), multiple sequence alignment (Sec. 2.2), alphabet reduction (Sec. 2.3), statistical model of sequence motif (Sec. 2.4), sequence scoring using a position weight matrix (Sec. 2.5). Finally, in Sec. 2.6, we introduce the pipeline to predict $\gamma-$secretase using the aforementioned bioinformatics tools.

2.1 Data Preparation

Training motif model for $\gamma-$secretase substrates requires proteins cleaved by $\gamma-$secretase as input. As there are only three experimentally verified substrates of $\gamma-$secretase, β amyloid protein precursor-like (Appl), Notch and Cadherin-N (CadN), we need to expand the training set by using homologous proteins of these substrates.

Since all substrates of $\gamma-$secretase are type I transmembrane protein, we first

use TMHMM [19], a hidden Markov model [20] based method to predict transmembrane regions within protein sequences and select transmembrane proteins from all *Drosophila* proteins. In total, we have found 673 transmembrane proteins from TMHMM prediction.

Next, we ran HHpred [21], a HMM-HMM comparison method that uses HMMs to encode protein evolutionary constraints, with human APP, human Notch, *Drosophila* APPL and *Drosophila* Notch as input to detect their homologous proteins amongst those 673 transmembrane proteins. This procedure results in a list of 112 proteins. Then we used the 112 proteins, together with the three known γ-secretase substrates, as a case data set, and the rest of the 673 proteins as a control to train the motif model.

2.2 Multiple Sequence Alignment

As mentioned in the previous chapter, there are two ways to train motif models: one is based on homologous information, where we find the consensus region of homologous proteins, another is the EM based algorithm, where given a list of unaligned sequences, we predict the site of the motif in each sequence and simultaneously train the motif model using the predicted sites. Since here we first chose the homologous based motif training method, selecting an effective method for multiple sequence alignment is an important step in our pipeline. Later in Sec. 3.3 we compare to the EM based algorithm and other methods to demonstrate our favourable performance.

Multiple sequence alignment assumes the input query sequences descended from a common ancestor and aligns them together through detecting homologus regions. The most simple way is through dynamic programming, which is computationally infeasible as it is NP-complete [22]. There are also other approaches, such as progressive alignment construction [23], iterative methods [24], consensus methods [25] and hidden Markov model [24], all of which cannot guarantee an optimal alignment from a theoretical perspective, but usually have good performance in real world data analysis.

Here we used mafft [26] for multiple sequence alignment. It is a multiple sequence alignment program that uses fast Fourier transform (FFT) to discover homologous segments and then builds up multiple sequence alignments through pairwise alignment from the most similar to the most different sequences. It is based on the observation that fast Fourier transform is a very efficient method to detect sequence similarities [27, 28].

2.3 Alphabet Reduction

In order to generate multiple sequence alignments, as the amino acid alphabet of proteins is too large, which significantly influences the performance of multiple sequence alignment, we used an alphabet shrinkage method to shrink the alphabet size to four letters [15]. There are many strategies to shrink alphabets based on biophysical similarity of amino acids.

In Bacardit et al. [15], the authors changed this problem into an optimization one that they solve by clustering amino acids by maximizing the objective function:

$$I(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

where for each instance in the database, x corresponds to a string encoding the "feature" (either contact number [29] or solvent accessibility [30]) of this instance and y denotes the cluster index this instance belongs to. All probabilities are calculated from their occurrence in the database. Based on maximizing the mutual information strategy, the authors [15] also made a few extensions beyond that. One is called the "robust mutual information" strategy, where they add a shuffling process to the original mutual information computation to avoid overestimation of small sample size. Another is called the "dual robust mutual information" strategy, which is developed based on a "robust mutual information" strategy that assigns different alphabet reduction strategy between target residues and non-target residues (see Fig. 1 and Fig. 2 of [15]).

From the experimental performance (which will be further discussed in the results chapter), we found the "dual robust mutual information" strategy with solvent accessibility features has the best performance, and choose that strategy to reduce the amino acid into a four-letter alphabet.

2.4 Position Weight Matrix for Sequence Motif Model

Position weight matrices (PWM) are commonly used to model sequence motifs, based on the assumption that a protein would interact with another biomolecule by recognizing a k-mer, a short continuous sequence fragment of length k, within the sequence.

PWMs match the $\gamma-$secretase case as the cleavage site of $\gamma-$secretase is within the transmembrane region of protein. In these regions, most transmembrane proteins would unfold into a more linear structure: $\alpha-$helix [31], for $\gamma-$secretase cleavage. Therefore, a k-mer model would be suitable to model $\gamma-$secretase cleavage considering its linear nature.

From a mathematical perspective, PWMs use matrix to model occurrence probabilities of amino acids in each position of the k-mer, where each row in the matrix corresponds to an amino acid and each column corresponds to a position in the k-mer. In this way, value of each cell $M_{l,j}$ in the matrix M represents the probability of occurrence of amino acid l at k-mer's position j, which is calculate as:

$$M_{l,j} = \frac{1}{N} \sum_{i=1}^{N} I(X_{i,j} = l),$$

here l is the amino acid of the row, $X_{i,j}$ is the letter of i-th sequence at j-th position of the k-mer in sequence, all indices are 1-indexed.

2.5 Scoring Motif Match Using Log-likelihood

Given the PWM of motif, M, we calculate the score of sequence s using the following formula:

$$score = \max_i \sum_{j=1}^{k} \log M_{s[i+j-1],j}.$$

Here k is the length of k-mer, $s[i]$ is the $i-$th amino acid in sequence s. The intuition is that if we consider a PWM as a generative model for a k-mer, we need to select the k-mer that has the highest generative probability in the matrix model. When the motifs for both case (M^{case}) and control (M^{ctrl}) data are provided, we can calculate the score using the log-likelihood ratio:

$$score = \max_i \sum_{j=1}^{k} \log \frac{M^{case}_{s[i+j-1],j}}{M^{ctrl}_{s[i+j-1],j}}.$$

2.6 Pipeline for $\gamma-$secretase Substrate Prediction

See Fig. 2-1 for an outline of the pipeline we used to predict $\gamma-$secretase substrates.

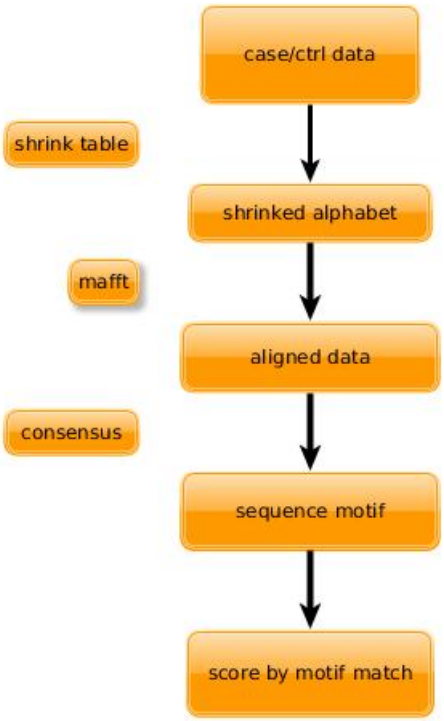


Figure 2-1: Pipeline for prediction of $\gamma-$secretase cleavage substrates

First, for all the 112 predicted substrates of $\gamma-$secretase, as the cleavage site of

$\gamma-$secretase is within the transmembrane region of the protein, we extracted transmembrane regions (predicted by TMHMM) with flanking regions on each side as input sequence for alignment, here we tested different flanking region length and found flanking regions with 50 amino acids on each side has the best performance. We also added regions extracted from the three experimentally verified $\gamma-$secretase substrates to the alignment input. When there are multiple regions of the protein predicted as transmembrane regions, we just consider all of them.

Then we used the alphabet reduction strategy described in Sec. 2.3 to reduce the original amino acid alphabet into a four-letter alphabet and run mafft on the sequence translated to the reduced alphabet. We also did the same thing on the control.

After a multiple sequence alignment has been constructed using mafft, for each protein in the training set, we checked the quality of the alignment to the three experimentally verified substrates and discarded the proteins with poor alignment. In particular, we manually selected a region of Appl such that the local alignment with $> 90\%$ sequence identity was kept to generate the motif.

Next, as the three experimentally verified substrates are also candidates in the multiple sequence alignment, for each of the three substrates, we constructed a PWM by first manually selecting a region in that substrate and using sequences aligned to that region to build a PWM. The coordinates of the regions we selected to build the PWM are listed in Table 2.1. Here for each substrate we generated motifs using both case and control sequence sets.

protein symbol	isoform id	start	end
Appl	FBpp0070104	819	842
Notch	FBpp0070483	1748	1771
CadN	FBpp0080569	2923	2946

Table 2.1: Coordinates of the alignment region used for generating $\gamma-$secretase substrate motif relative to three experimentally verified substrates. Notice they are all aligned with each other and have high alignment coverage.

After we have gotten the PWM of the substrate motif, we used the PWM to predict scores for $\gamma-$secretase cleavage using the method of Sec. 2.5. Both using only the case motif and using the case and control motifs together are considered to

compare performance.

2.7 Empirical Reasons for Choosing Alignment Based Method

We choose the alignment based method relative to non-alignment based method for the empirical reason that:

- training data is composed of homologous proteins, thus alignment based methods are able to fully utilize homology information as compared to non-alignment based methods.

In the results chapter, we support this empirical insight with computational validation on benchmark data set.

Chapter 3

Results

3.1 Motifs Generated by Multiple Sequence Alignment

Once we have aligned case sequences, we need to select regions for each experimentally verified substrate and use the sequences aligned to that region to generate a substrate motif.

First, for each experimentally validated substrate, we selected a flanking region of 20 amino acids on each side of the cleavage site and used the sequences aligned to this region to generate the motif. The reason for choosing flanking length as 20 is that the regions we chose to build up PWM (Table 2.1) are all within 20 amino acids from the cleavage site, so we picked flanking regions of 20 for visualization. The motif in Fig. 3-1 was generated by weblogo [32].

We also used the same approach to generate motifs using control data (Fig. 3-2), for consistency we chose the same coordinate as cases. By comparing cases and controls, we have found several differences:

- motifs for case sequences look more similar, due mostly to better alignment quality so that selected regions of the three substrates are aligned together;
- amino acid distribution for the control motif tend to be more uniform than

case, which means case sequences tend to be more homologous than control sequences.

As the number of control sequences is much larger than case, and a larger number of alignment candidates could significantly reduce the alignment quality, here we also randomly selected control sequences with equal size to the case and constructed the motif based on the sampled set (Fig. 3-3).

We also summarized the alignment coverage: percentage of amino acids aligned to those regions rather than gaps (red line in Fig. 3-4) to measure the motif quality. Alignment coverage of the control set is depicted by the green line.

Based on both case motif and alignment coverage data, we further shrink the motif into the red rectangular region to build motif for prediction, it is based on two reasons:

- red rectangular region of all three motifs are aligned together in the multiple sequence alignment;
- the alignment coverage of the region in the dark red rectangular is very high (red line in Fig. 3-4).

From motif shown in Fig. 3-1, it is interesting to see that the motif contains a segment of medium-size hydrophobic residues, such as Leucine (L) and Valine (V), and a cluster of positively charged residues, such as Arginine (R). This motif is consistent with a hypothetical model described in a recent manuscript by Bai et al. [33], which provides biological evidence to support our computational motif discovery.

3.2 Motif Prediction Result

After the motif finding procedure, we also tested the predictive performance on two benchmark data sets we collected. Details of how the two data sets are generated is listed here:

- first benchmark: the 112 proteins used in the training case motif as positive, all predicted type I transmembrane proteins as negative;

- second benchmark: *Drosophila* proteins predicted to be orthologous to human $\gamma-$secretase substrates (DIOPT [34] as prediction tool, with DIOPT score at least 1) as positive, all predicted type I transmembrane proteins as negative.

After the benchmark data was generated, we used motifs generated in Sec. 3.1 to make predictions with three methods:

- likelihood method: calculate likelihood of each sequence using motif from case data;

- likelihood ratio method: calculate likelihood ratio of each sequence using motif from both case and control data;

- likelihood ratio (shrink) method: calculate likelihood ratio of each sequence using motifs from case data and randomly sampled control data.

Method		AUC value
first benchmark	likelihood	0.80
	likelihood ratio	0.81
	likelihood ratio (shrink)	0.82
second benchmark	likelihood	0.69
	likelihood ratio	0.67
	likelihood ratio (shrink)	0.71

Table 3.1: AUC value of method on benchmark dataset

Results indicate that when using case data, the performance of the three models looks nearly identical (Table 3.1), however, this could be caused by overfitting of benchmark. When transformed into a completely new data set (benchmark 2), the likelihood ratio (shrink) has the best performance, and thus we used this method to rank all type I transmembrane proteins we discovered (Table 3.2). And we would use the top 10 transmembrane proteins for experimental validation.

Rank			fbgnid	Gene Symbol
Motif	APPL	Notch		
1	17	2	FBgn0011592	fra
2			FBgn0015609	CadN
3			FBgn0036202	CG6024
4			FBgn0030603	CG5541
5			FBgn0051072	Lerp
6	1		FBgn0000108	Appl
7			FBgn0261574	kug
8			FBgn0040256	Ugt86Dd
9			FBgn0004370	Ptp10D
10			FBgn0005631	robo1

Table 3.2: Prediction result of type I transmembrane proteins: "Rank" means the rank of score using motif model, "APPL" means rank of homologous protein of Appl, same for "Notch", only top 10 is shown.

Table 3.2 contains some of the top ranked type I transmembrane proteins. Due to space limitions, here we just show the top 10 proteins in the prediction list, the rest are in Supplement (Table A.1). From the results we see both Appl (ranking 6) and CadN (ranking 2) have very high predicted ranks, but surprisingly Notch is not in the top list (the rank is 13); only a protein highly homologous to Notch is in the top list. This is a reasonable prediction error.

3.3 Comparison with Other Methods

3.3.1 Comparison with Non-alignment Based Method

Here we also compared with a non-alignment based method. As mentioned in the introduction chapter, there are two ways to find sequence motif, one is by using multiple sequence alignment and then build a position weight matrix based on alignment result, which we have done; another is by using an EM based algorithm to predict the binding site of each sequence and construct a position weight matrix through occurrence of amino acids at the predicted sites.

There are two additional requirements if we want to use evolutionary based methods rather than the straightforward EM-based algorithm:

- input sequences are homologous proteins;
- we are certain about where the binding site is for at least one protein in the multiple sequence alignment.

Fortunately the provided dataset happens to satisfy these requirements, so empirically evolutionary based methods should have better performance. Here for comparison I used the EM based algorithm MEME [17] to generate the motif.

To run MEME, we set the maximum number of motifs generated to 5, As MEME selects a subset of sequences to generate a motif, we set the minimum number of sequences in generating this motif to 200. As DREME does not support protein sequences, here we just consider MEME. See Fig. 3-5 for motifs generated from case sequences and Fig. 3-6 for motifs from control sequences.

We can see the first discovered motif using case data (Fig. 3-5a), as compared to control data (Fig. 3-6), matches our empirical evaluation best ("L" enriched), so we used it as the motif from case set. For the control motif, we chose Fig. 3-6a. After motifs were generated, we tested their performance using the aforementioned benchmark, see Table 3.3 for results. We can see the MEME suite using the likelihood method has the best performance in this test.

Method		AUC value
first benchmark	likelihood	0.71
	likelihood ratio	0.66
second benchmark	likelihood	0.62
	likelihood ratio	0.61

Table 3.3: AUC value for non-alignment based method on benchmark dataset using MEME motif

By comparing Tables 3.1 and 3.3, we observe the alignment based method significantly outperforms the non-alignment based method, which also matches our empirical observations.

The superior performance is likely due to that case data are mainly generated from homology search, and homology based methods can fully use the evolutionary information to align cleavage sites of candidates for more accurate inference of the

motif. In contrast, the EM based method cannot fully utilize this information, but could be more effective when the case set is not comprised of homologous proteins.

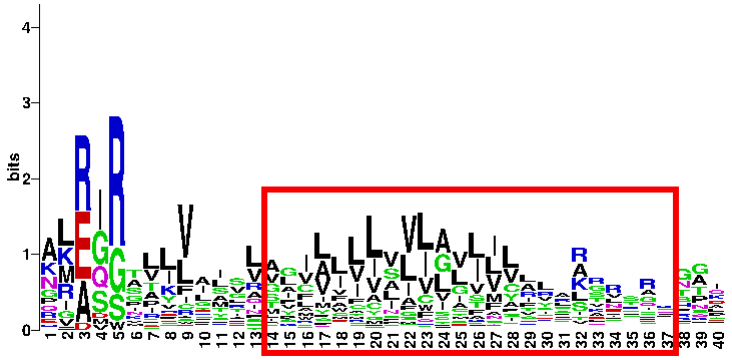
3.3.2 Comparison with Alignment Based Method without Alphabet Reduction

Here we also examine the affect of alphabet reduction. First we generated case motifs using multiple sequence alignment without alphabet reduction (Fig. 3-8) and used this motif for prediction (Table 3.4).

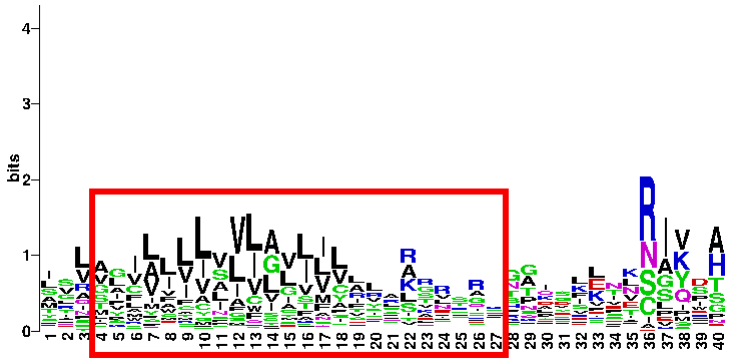
experiment		AUC value
case data	case motif only	0.68
	case and control motif	0.74
	case and control motif (shrink)	0.63
experiment data	case motif only	0.67
	case and control motif	0.66
	case and control motif (shrink)	0.67

Table 3.4: AUC value of method without alphabet reduction procedure on benchmark dataset

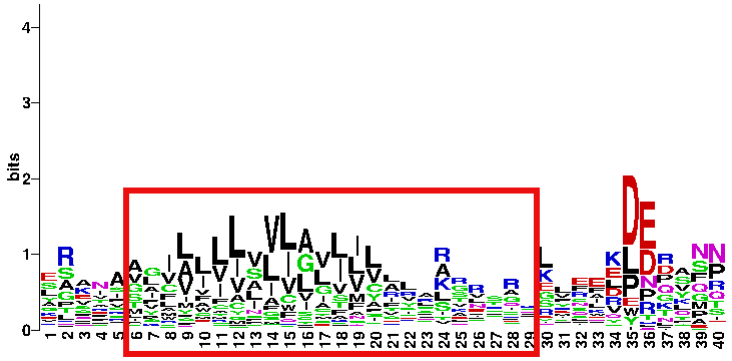
From the AUC result, as well as the motif picture, we see that alphabet reduction significantly improves alignment quality and thus creates a more accurate motif model for prediction.



(a) Appl

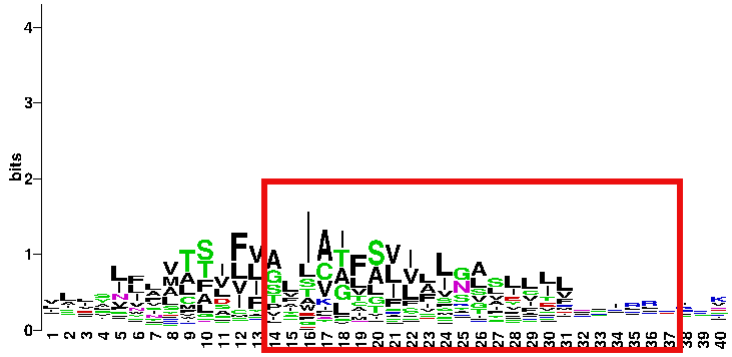


(b) Notch

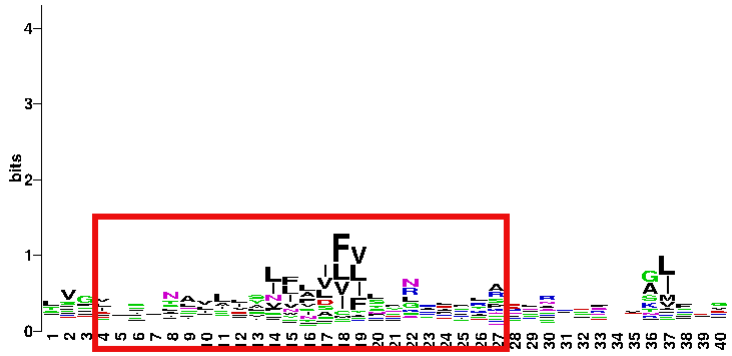


(c) CadN

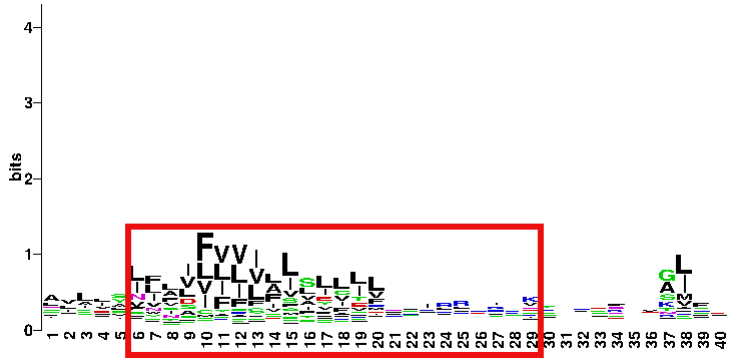
Figure 3-1: Motif of three experimentally verified substrates using case data; regions within red rectangle are the selected regions for position weight matrix



(a) Appl

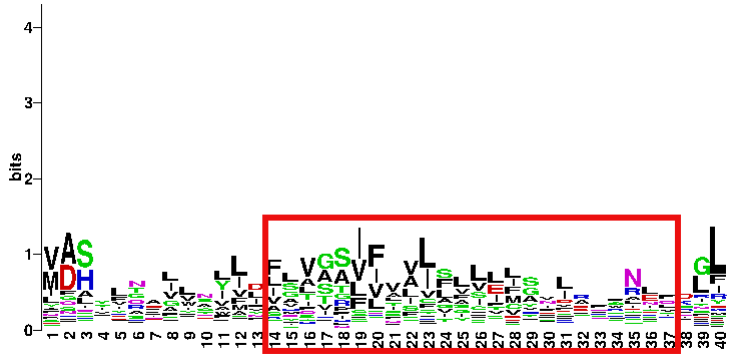


(b) Notch

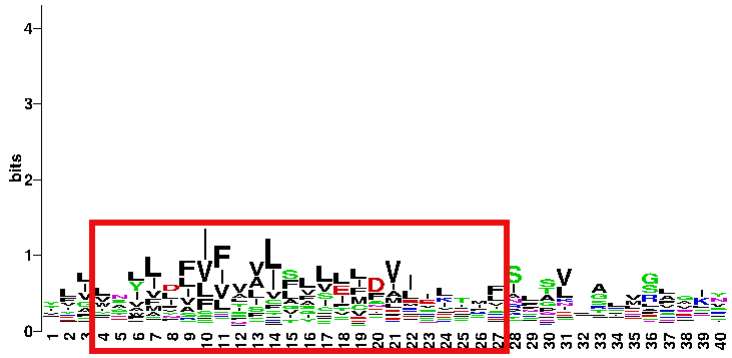


(c) CadN

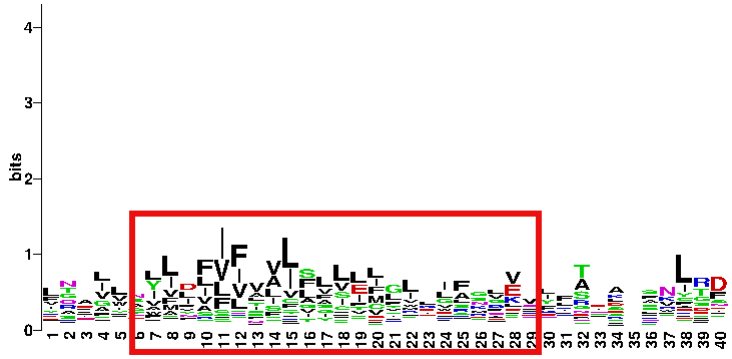
Figure 3-2: Motif of three experimentally verified substrates using control data; regions within red rectangle are the selected regions for position weight matrix, for consistency we selected the same coordinates as case motifs.



(a) Appl

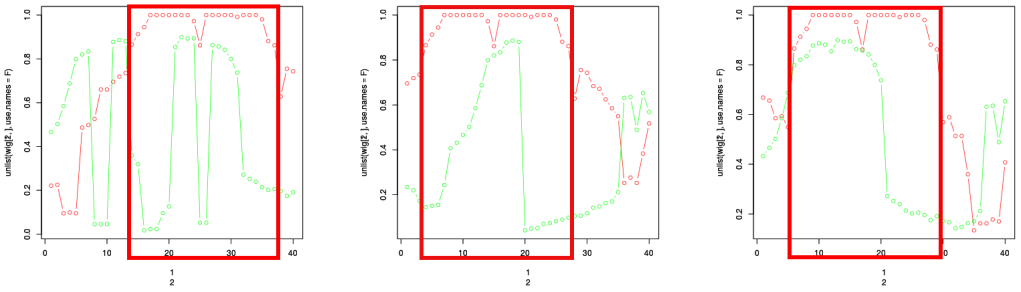


(b) Notch



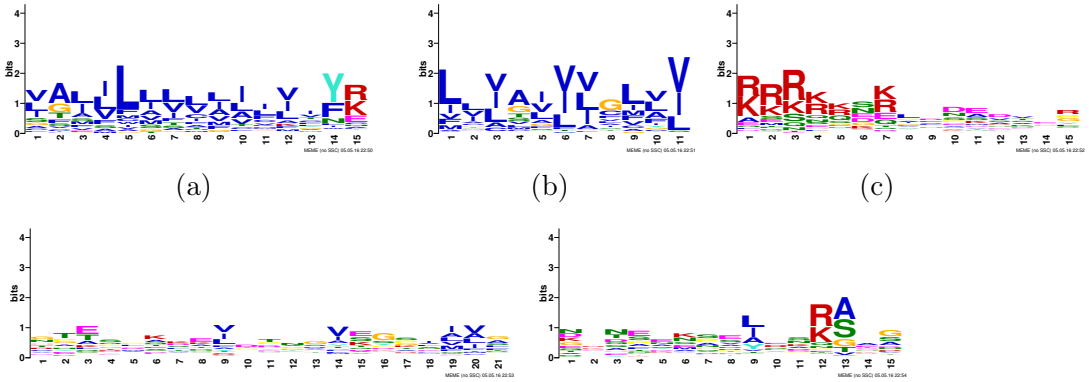
(c) CadN

Figure 3-3: Motif found for three experimentally verified substrates using (shrink) control data set of equal size to case set through random sampling; regions within red rectangle are the selected regions for position weight matrix.



(a) Appl (b) Notch (c) CadN

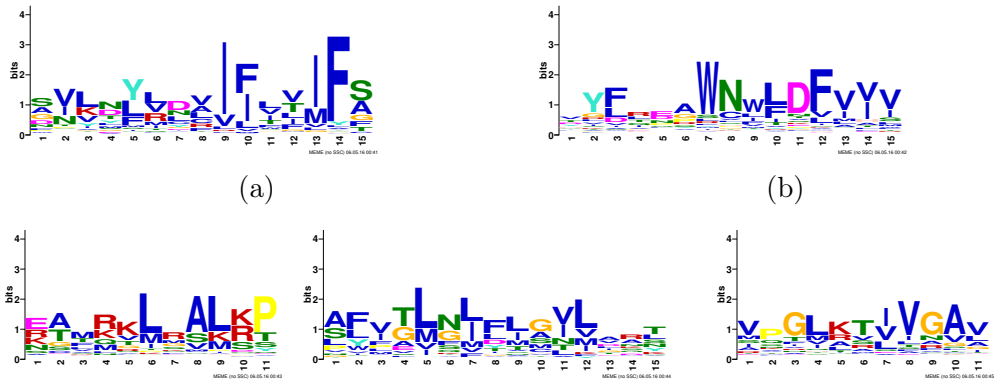
Figure 3-4: Alignment coverage for both case sequence alignment (red line) and control sequence alignment (green line), y-axis is the percentage of amino acid coverage for each position, x-axis is the same as the x-axises in previous motif logos; regions within red rectangle are the selected regions for position weight matrix.



(a) (b) (c)

(d) (e)

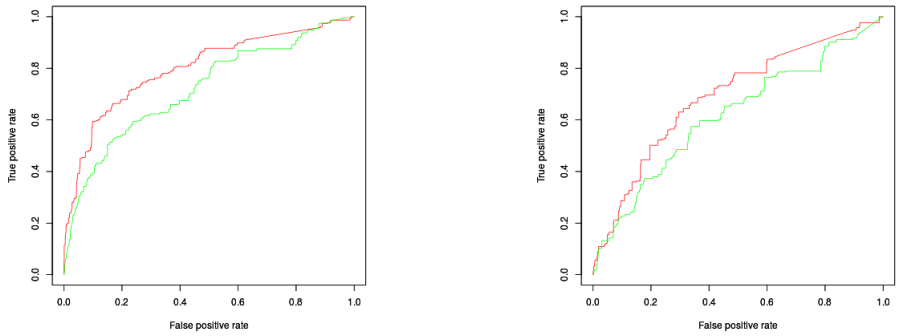
Figure 3-5: Non-alignment based motif found using MEME suite with case data



(a) (b)

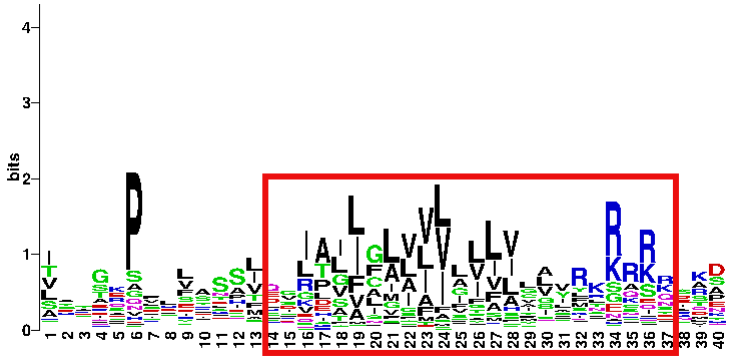
(c) (d) (e)

Figure 3-6: Non-alignment based motif found using MEME suite with control data

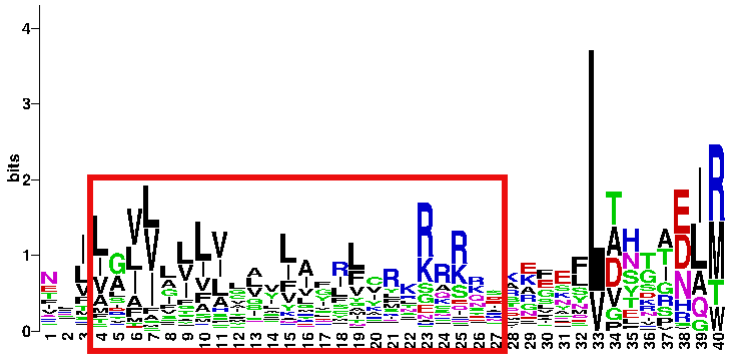


(a) ROC curve for training data (b) ROC curve for experimental data

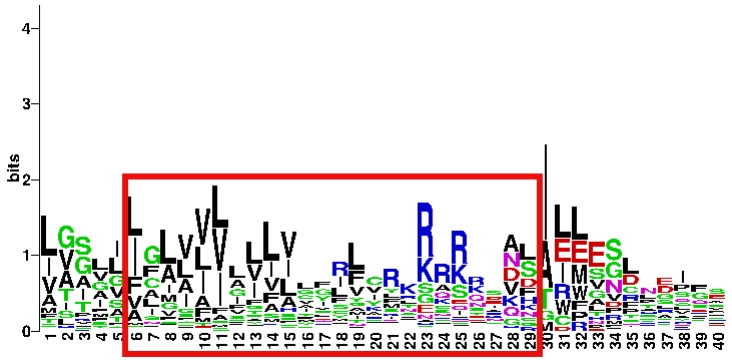
Figure 3-7: Comparison of ROC curve for non-alignment based methods using homologous motif and MEME motif; red line corresponds to using homologous motif, green line to using MEME motif



(a) Appl



(b) Notch



(c) CadN

Figure 3-8: Motif of three experimentally verified substrates using case data without alphabet reduction; regions within red rectangle are the selected regions for position weight matrix

Chapter 4

Conclusion

In this project, we have proposed a computational method to predict substrates of $\gamma-$secretase based on motifs devised through a multiple sequence alignment pipeline. Importantly, a reduced amino acid alphabet was a critical component. We also designed several experiments to compare its performance with other popular methods.

Overall, we expect predictions resulting from this computational method to be very helpful in testing two hypotheses of Alzheimer's disease: presenilin hypothesis and amyloid hypothesis. This experiment also shows that sequence motif based methods are be very helpful in predicting protein interactions within transmembrane proteins, which could be generalized in principle to other transmembrane binding problems.

Appendix A

Tables

Rank			fbgnid	Gene Symbol
Motif	**APPL**	**Notch**		
1	17	2	FBgn0011592	fra
2			FBgn0015609	CadN
3			FBgn0036202	CG6024
4			FBgn0030603	CG5541
5			FBgn0051072	Lerp
6	1		FBgn0000108	Appl
7			FBgn0261574	kug
8			FBgn0040256	Ugt86Dd
9			FBgn0004370	Ptp10D
10			FBgn0005631	robo1
11			FBgn0030310	PGRP-SA
12			FBgn0261262	CG42613
13	16	1	FBgn0004647	N
14			FBgn0262018	CadN2
15			FBgn0030090	fend
16			FBgn0030769	CG13012
17			FBgn0013997	Nrx-IV

18			FBgn0038868	CG5862
19			FBgn0011283	Obp28a
20			FBgn0032235	CG5096
21			FBgn0038028	CG10035
22			FBgn0032713	CG17323
23			FBgn0034718	wdp
24	5	23	FBgn0261822	Bsg
25			FBgn0028400	Syt4
26			FBgn0034797	nahoda
27	22	4	FBgn0038638	CG7702
28	15	22	FBgn0010473	tutl
29			FBgn0011204	cue
30			FBgn0031735	CG11029
31			FBgn0037007	CG5059
32			FBgn0265182	CG44247
33			FBgn0031305	Iris
34			FBgn0031879	uif
35			FBgn0035543	CG15020
36			FBgn0002543	robo2
37			FBgn0040388	boi
38			FBgn0016061	side
39			FBgn0051036	CG31036
40			FBgn0015380	drl
41			FBgn0034066	CG8397
42			FBgn0033403	CG13739
43			FBgn0032856	CG16798
44			FBgn0036509	CG7739
45			FBgn0024245	dnt

46			FBgn0032851	CG13970
47			FBgn0035833	CG7565
48			FBgn0263219	Dscam4
49			FBgn0029922	CG14431
50			FBgn0010395	Itgbn
51			FBgn0030260	CG1537
52			FBgn0039265	CG11790
53			FBgn0039140	Miro
54			FBgn0030703	MSBP
55			FBgn0001083	fw
56		30	FBgn0085400	CG34371
57			FBgn0036587	CG4950
58			FBgn0034642	CG15674
59			FBgn0036146	CG14141
60			FBgn0263830	CG40486
61			FBgn0004242	Syt1
62			FBgn0259245	DIP-beta
63			FBgn0034541	CG13437
64			FBgn0264975	Nrg
65			FBgn0029866	CG3842
66			FBgn0051262	CG31262
67			FBgn0263256	CG43394
68			FBgn0002306	sas
69			FBgn0031554	CG15418
70			FBgn0003716	tkv
71			FBgn0033777	CG17574
72			FBgn0085409	CG34380
73			FBgn0030992	CG33253

74			FBgn0028482	bdl
75			FBgn0004369	Ptp99A
76			FBgn0003366	sev
77			FBgn0011300	babo
79			FBgn0000094	Anp
80			FBgn0000636	Fas3
81			FBgn0033880	CG6553
82			FBgn0004511	dy
83			FBgn0035483	Mul1
84			FBgn0035040	CG4741
85			FBgn0003285	rst
86			FBgn0028369	kirre
88			FBgn0026314	Ugt35b
89			FBgn0038871	CG3337
90			FBgn0035043	CG4781
91			FBgn0053087	LRP1
92			FBgn0051068	CG31068
93			FBgn0085292	CG34263
94			FBgn0262482	CG43072
95			FBgn0036928	Tom20
96			FBgn0261260	mgl
97			FBgn0031571	bark
98			FBgn0264739	CG43997
99			FBgn0034031	CG12963
100			FBgn0035429	CG12017
101			FBgn0039086	CG16732
102			FBgn0044809	TotZ
103			FBgn0250832	Dup99B

104			FBgn0260223	CG42497
105			FBgn0261277	rtv
106			FBgn0035444	CG12012
107			FBgn0025686	Amnionless
108			FBgn0010482	l(2)01289
109			FBgn0039521	CG5402
110			FBgn0066365	dyl
111			FBgn0031560	CG16713
112			FBgn0085423	CG34394
113			FBgn0266757	mfr
114			FBgn0016047	nompA
115			FBgn0031564	CG2816
116			FBgn0261358	CG42635
117			FBgn0003391	shg
118			FBgn0039928	Cals
119			FBgn0034005	ItgaPS4
120			FBgn0086604	CG12484
121			FBgn0033814	CG4670
122			FBgn0015622	Cnx99A
123			FBgn0030847	CG12991
124			FBgn0029669	CG13021
125			FBgn0033020	COX4L
126			FBgn0260954	CG42586
127			FBgn0028537	CG31775
128			FBgn0051431	CG31431
129			FBgn0026754	Ugt37c1
130			FBgn0052141	CG32141
131			FBgn0031294	IA-2

132			FBgn0030993	Mec2
133			FBgn0053144	CG33144
134			FBgn0038809	CG16953
135			FBgn0034540	Lrt
136			FBgn0051004	mesh
137			FBgn0052792	ppk8
138			FBgn0250821	CG14644
139			FBgn0004648	svr
140			FBgn0040250	Ugt86Dj
141	14	8	FBgn0000463	Dl
142			FBgn0085377	CG34348
143			FBgn0038610	CG7675
144			FBgn0028872	CG18095
145			FBgn0264303	CG43781
146			FBgn0262515	VhaAC45
147			FBgn0033943	CG12869
148			FBgn0085199	CG34170
149			FBgn0264302	CG43780
150			FBgn0037030	CG3288
151			FBgn0032860	CG15130
152			FBgn0083975	Nlg4
153			FBgn0083963	Nlg3
154			FBgn0038156	CG14372
155			FBgn0042179	CG18869
156			FBgn0004591	Eig71Ed
157			FBgn0263083	CG43351
158			FBgn0004197	Ser
159			FBgn0037110	ORMDL

160			FBgn0038127	CG8476
161			FBgn0043903	dome
162			FBgn0050373	CG30373
163			FBgn0031548	CG8852
164			FBgn0262870	axo
165			FBgn0259110	mmd
166			FBgn0019985	mGluR
167			FBgn0259238	CG42336
168			FBgn0005672	spi
169			FBgn0031969	pes
170			FBgn0040261	Ugt36Bb
171			FBgn0037736	CG12950
172			FBgn0259190	Ir7d
173			FBgn0033519	CG11825
174			FBgn0034476	Toll-7
175			FBgn0053523	Vap-33B
176			FBgn0261801	CG42747
177			FBgn0036380	CG8757
178			FBgn0261975	CG42806
179			FBgn0260230	CG42504
180			FBgn0037416	Osi9
181			FBgn0085279	CG34250
182			FBgn0040636	CG13255
183			FBgn0046875	Obp83g
184			FBgn0264255	para
185	24	24	FBgn0262509	nrm
186			FBgn0051676	CG31676
187			FBgn0034468	Obp56a

188			FBgn0027074	CG17324
189			FBgn0034605	CG15661
190			FBgn0039234	nct
191			FBgn0051665	wry
192			FBgn0037411	Osi3
193			FBgn0039852	nyo
194			FBgn0261509	haf
195			FBgn0262730	CG14446
196			FBgn0052667	ssp7
197			FBgn0263249	CG43392
198			FBgn0002577	m
199			FBgn0262788	CG43169
200			FBgn0085485	CG34456
201			FBgn0260768	CG42566
202			FBgn0037908	dpr5
203		44	FBgn0066101	LpR1
204			FBgn0004657	mys
205			FBgn0029687	Vap-33A
206			FBgn0083991	CG34155
207			FBgn0003377	Sgs7
208			FBgn0264000	GluRIB
209			FBgn0034671	CG13494
210			FBgn0052240	CG32240
211			FBgn0051496	CG31496
212			FBgn0035020	CG13585
213			FBgn0085398	ppk9
214			FBgn0037134	CG7407
215			FBgn0037421	CG15594

216			FBgn0037963	Cad87A
217			FBgn0032484	kek4
218			FBgn0032095	Toll-4
219			FBgn0266801	CG45263
220			FBgn0085322	CG34293
221			FBgn0001137	grk
222			FBgn0259896	NimC1
223			FBgn0039723	CG15522
224			FBgn0004118	nAChRbeta2
225			FBgn0042180	CG18870
226			FBgn0002873	mud
227			FBgn0024983	CG4293
228			FBgn0030319	CG2533
229			FBgn0050495	CG30495
230			FBgn0264077	Cnx14D
231		10	FBgn0259685	crb
232			FBgn0033313	Cirl
233			FBgn0039704	neo
234			FBgn0053143	CG33143
235			FBgn0030440	CG15719
236	10		FBgn0010452	trn
237			FBgn0024189	sns
238			FBgn0261674	CG42709
239			FBgn0039709	Cad99C
240			FBgn0259717	CG42371
241	3	3	FBgn0027594	drpr
242			FBgn0028430	He
243			FBgn0029868	ND-B16.6

244			FBgn0053493	CG33493
245			FBgn0050125	Ir56a
246			FBgn0261538	CG42662
247			FBgn0250845	CG1288
248			FBgn0034730	ppk12
249			FBgn0000119	arr
250			FBgn0259202	CG42306
251			FBgn0020521	pio
252			FBgn0035976	PGRP-LC
253			FBgn0036488	CG6878
254			FBgn0039087	CG10168
255			FBgn0004619	GluRIA
256			FBgn0051092	LpR2
257			FBgn0030174	CG15312
258			FBgn0040931	CG9034
259			FBgn0011016	SsRbeta
260			FBgn0261567	CG42681
261			FBgn0262794	CG43175
262			FBgn0028939	NimC2
263	4		FBgn0030001	cyr
264			FBgn0040212	Dhap-at
265			FBgn0003984	vn
266			FBgn0037406	Osi1
267			FBgn0061492	loj
268			FBgn0259991	CG42488
269			FBgn0032006	Pvr
270			FBgn0031275	GABA-B-R3
271			FBgn0259185	Ir60b

272			FBgn0085339	CG34310
273			FBgn0031950	Herp
274			FBgn0083950	CG34114
275		46	FBgn0034602	Lapsyn
276			FBgn0010309	pigeon
277			FBgn0034072	Dg
278			FBgn0000152	Axs
279			FBgn0053196	dpy
280			FBgn0038098	CG7381
281			FBgn0265266	CG13639
282			FBgn0034050	CG8297
283			FBgn0030706	CG8909
284			FBgn0035785	ppk26
285			FBgn0040743	CG15919
286			FBgn0036173	CG7394
287			FBgn0038083	CG5999
288			FBgn0014868	Ost48
289			FBgn0000497	ds
290			FBgn0264908	pHCl
291			FBgn0034206	CG18469
292			FBgn0037016	CG13252
293			FBgn0035699	CG13300
294			FBgn0051360	CG31360
295			FBgn0032434	CG5421
296			FBgn0051913	CG31913
297		29	FBgn0029082	hbs
298			FBgn0265140	Meltrin
299			FBgn0260775	DnaJ-60

300			FBgn0053702	CG33702
301			FBgn0031049	Sec61gamma
302			FBgn0260452	CG13984
303			FBgn0051105	ppk22
304			FBgn0039527	CG5639
305			FBgn0261566	CG42680
306			FBgn0039810	CG15549
307			FBgn0037012	Rcd2
308			FBgn0040551	CG11686
309			FBgn0051323	CG31323
310			FBgn0031478	CG8814
311			FBgn0034737	CG11362
312			FBgn0037413	Osi5
313			FBgn0052313	CG32313
314			FBgn0031164	CG1724
315			FBgn0034578	CG15653
316			FBgn0038460	CG18622
317			FBgn0026619	Taz
318			FBgn0033691	CG8860
319			FBgn0035258	CG13931
320			FBgn0031190	CG12576
321			FBgn0010638	Sec61beta
322			FBgn0262508	CG43078
323			FBgn0034389	Mctp
324			FBgn0050411	CG30411
325			FBgn0039177	CG13611
326			FBgn0032752	CG10702
327			FBgn0085207	CG34178

328			FBgn0259916	CG42445
329			FBgn0038682	CG5835
330			FBgn0052283	Drsl3
331			FBgn0050381	CG30381
332			FBgn0052179	Krn
333			FBgn0261053	Cad86C
334			FBgn0025820	JTBR
335			FBgn0037419	Osi12
336			FBgn0261514	NimA
337			FBgn0039528	dsd
338			FBgn0039160	CG5510
339			FBgn0250876	Sema-5c
340			FBgn0036698	CG7724
341			FBgn0033405	CG13954
342			FBgn0085274	CG34245
343			FBgn0265187	CG44252
344			FBgn0051146	Nlg1
345			FBgn0038761	CG17190
346			FBgn0262686	CG43156
347			FBgn0262867	Ptr
348			FBgn0085320	CG34291
349			FBgn0004598	Fur2
350			FBgn0034545	CG13438
351			FBgn0039942	CG17163
352			FBgn0013272	Gp150
353			FBgn0037238	CG1090
354			FBgn0085201	CG34172
355			FBgn0015770	MstProx

356			FBgn0000277	CecA2
357			FBgn0000276	CecA1
358			FBgn0266124	ghi
359			FBgn0031505	ND-B14.5B
360			FBgn0051220	CG31220
361			FBgn0062442	Cisd2
362			FBgn0034717	CG5819
363			FBgn0039321	CG10550
364			FBgn0040324	Ephrin
365			FBgn0028327	l(1)G0320
366			FBgn0263031	CG43326
367			FBgn0029131	Debcl
368			FBgn0034498	CG16868
369			FBgn0037958	CG6962
370			FBgn0034554	CG15227
371			FBgn0259932	CG42455
372			FBgn0051774	fred
373			FBgn0025936	Eph
374			FBgn0034880	ItgaPS5
375			FBgn0011596	fzo
376			FBgn0033168	CG11145
377			FBgn0036145	CG7607
378			FBgn0040571	CG17193
379			FBgn0040251	Ugt86Di
380			FBgn0039431	plum
381			FBgn0034083	lbk
382			FBgn0052230	ND-MLRQ
383			FBgn0052450	CG32450

384			FBgn0261836	Msp300
385			FBgn0036008	CG3408
386			FBgn0030941	wgn
387		34	FBgn0266420	Ote
388			FBgn0003731	Egfr
389			FBgn0053310	CG33310
390			FBgn0262530	CG43084
391			FBgn0053531	Ddr
392			FBgn0042119	Cpr65Au
393			FBgn0260231	CG42505
394			FBgn0037796	CG12814
395			FBgn0263997	CG43740
396			FBgn0032013	Scgalpha
397			FBgn0000635	Fas2
398			FBgn0036286	CG10616
399			FBgn0053003	CG33003
400			FBgn0051002	CG31002
401			FBgn0046294	CG12699
402			FBgn0262467	Scox
403			FBgn0028475	Hrd3
404			FBgn0034270	CG6401
405			FBgn0052280	CG32280
406			FBgn0262838	CG43202
407			FBgn0036715	Cad74A
408			FBgn0265416	Neto
409			FBgn0035971	CG4477
410			FBgn0036670	CG13029
411			FBgn0037151	CG7130

412			FBgn0262823	CG43194
413			FBgn0035471	Sc2
414			FBgn0264478	CG43886
415			FBgn0039068	CG13827
416			FBgn0036978	Toll-9
417			FBgn0267428	CG45781
418			FBgn0043792	CG30427
419			FBgn0053481	dpr7
420			FBgn0037553	CG18249
421			FBgn0259677	CG42346
422	21		FBgn0026566	CG1307
423			FBgn0030991	CG7453
424			FBgn0032217	CG4972
425			FBgn0039811	CG15550
426			FBgn0259204	CG42308
427			FBgn0031981	CG7466
428			FBgn0031887	CG11289
429			FBgn0050438	CG30438
430			FBgn0267488	Mcr
431			FBgn0259821	CG42402
432			FBgn0033593	Listericin
433			FBgn0040491	Buffy
434		36	FBgn0001987	Gli
435			FBgn0040773	COX7C
436			FBgn0038886	CG6475
437			FBgn0034140	Lst
438			FBgn0262473	Tl
439			FBgn0037133	CG7370

440			FBgn0035290	dsb
441			FBgn0001075	ft
442			FBgn0266084	Fhos
443			FBgn0266696	CG45186
444			FBgn0033703	CG13170
445			FBgn0004456	mew
446			FBgn0265296	Dscam2
447			FBgn0259992	CG42489
448			FBgn0040719	CG15357
449			FBgn0027073	CG4302
450			FBgn0045823	vsg
451			FBgn0031518	CG3277
452			FBgn0260234	Xport-B
453			FBgn0001989	ND-B17
454			FBgn0026756	Ugt37a1
455			FBgn0039677	ppk30
456	19	5	FBgn0243514	eater
457			FBgn0030670	Pis
458			FBgn0037719	bocks
459			FBgn0263046	CG43341
460			FBgn0262356	CG43054
461			FBgn0262567	CG43107
462			FBgn0021979	l(2)k09913
463			FBgn0031058	CG14227
464			FBgn0038639	CG7705
465			FBgn0035032	ATPsynF
466			FBgn0032233	dpr19
467			FBgn0040715	CG15386

468			FBgn0263086	CG43354
469			FBgn0036627	Gagr
470			FBgn0036690	Ilp8
471			FBgn0259950	CG42460
472			FBgn0028942	CG16852
473			FBgn0034568	CG3216
474	25	11	FBgn0031872	ihog
475			FBgn0261634	CG42717
476			FBgn0030868	CG12986
477			FBgn0041160	comm2
478			FBgn0039357	CG4743
479			FBgn0031080	CG12655
480			FBgn0025558	CG4101
481			FBgn0040514	CG17169
482			FBgn0262537	CG43091
483			FBgn0266580	Gp210
484			FBgn0029603	CG14053
485			FBgn0262790	CG43171
486			FBgn0036851	CG14082
487			FBgn0262791	CG43172
488			FBgn0029838	CG4666
489			FBgn0004055	uzip
490			FBgn0040091	Ugt58Fa
491			FBgn0034368	CG5482
492	8		FBgn0039969	Fis1
493			FBgn0261925	CG42792
494			FBgn0035346	CG1146
495			FBgn0040968	CG14933

496			FBgn0034639	CG15673
497			FBgn0050222	CG30222
498			FBgn0010415	Sdc
499			FBgn0243486	rdo
500			FBgn0034122	CG15711
501			FBgn0037530	EMC1
502			FBgn0262837	CG43201
503			FBgn0039031	CG17244
504			FBgn0265188	CG44253
505			FBgn0010548	Aldh-III
506			FBgn0259735	CG42389
507			FBgn0041097	robo3
508			FBgn0040805	CG12355
509			FBgn0052037	CG32037
510			FBgn0035094	CG9380
511			FBgn0259971	CG42481
512			FBgn0032900	CG14401
513			FBgn0003997	hid
514			FBgn0264561	Glg1
515			FBgn0261999	CG42817
516			FBgn0085399	CG34370
517			FBgn0052521	CG32521
518			FBgn0264297	CG43775
519			FBgn0036980	RhoBTB
520			FBgn0032474	DnaJ-H
521			FBgn0263971	CG43725
522			FBgn0038515	CG5823
523			FBgn0263761	CG43678

524			FBgn0263621	CG43630
525			FBgn0263247	CG43390
526			FBgn0263248	CG43391
527			FBgn0040832	CG8012
528			FBgn0261550	CG42668
529			FBgn0262840	CG43204
530			FBgn0262683	CG43153
531			FBgn0039172	Spase22-23
532			FBgn0032129	jp
533			FBgn0261991	CG42809
534			FBgn0036586	CG13070
535			FBgn0261697	tectonic
536			FBgn0028572	qtc
537			FBgn0040011	Slmap
538			FBgn0029728	CG2861
539			FBgn0032336	AstC
540			FBgn0038451	CG14893
541			FBgn0261984	Ire1
542			FBgn0053155	CG33155
543			FBgn0033645	CG13196
544			FBgn0050104	NT5E-2
545			FBgn0039666	Diedel
546			FBgn0023178	Pdf
547			FBgn0039356	CG5039
548			FBgn0039188	Golgin84
549			FBgn0051198	CG31198
550			FBgn0038751	CG4770
551			FBgn0038656	CG14294

552			FBgn0037828	tomboy20
553			FBgn0037679	CG8866
554			FBgn0040532	CG8369
555			FBgn0051787	CG31787
556			FBgn0028520	CG4891
557			FBgn0031849	CG11327
558			FBgn0031779	CG9175
559			FBgn0031737	obst-E
560			FBgn0037105	S1P
561			FBgn0040842	CG15212
562			FBgn0035880	CG17352
563			FBgn0000358	Cp19
564			FBgn0052069	CG32069
565			FBgn0036643	Syx8
566			FBgn0036938	CG14187
567			FBgn0030723	dpr18
568			FBgn0035518	CG15011
569			FBgn0029750	CG3323
570			FBgn0025645	CG3598
571			FBgn0037933	Ho
572			FBgn0040877	CG12994
573			FBgn0034172	CG6665
574			FBgn0050377	CG30377
575			FBgn0044047	Ilp6
576			FBgn0050355	CG30355
577			FBgn0039085	CG10170
578			FBgn0262808	CG43179
579			FBgn0262846	CG43210

580			FBgn0034129	CG15925
581			FBgn0024985	CG11448
582			FBgn0029128	tyn
583			FBgn0034861	CG9815
584			FBgn0036221	CG11588
585			FBgn0264389	opm
586			FBgn0264991	CG44142
587			FBgn0003328	scb
588			FBgn0261628	CG42711
589			FBgn0004649	yl
590			FBgn0037199	CG11137
591			FBgn0040899	CG17776
592			FBgn0051644	CG31644
593			FBgn0264089	sli
594			FBgn0038602	CG7126
595			FBgn0038631	CG7695
596			FBgn0264543	CG43922
597			FBgn0038071	Dtg
598			FBgn0033961	ND-B15
599			FBgn0035909	ergic53
600			FBgn0029696	CG15571
601			FBgn0028331	l(1)G0289
602			FBgn0035593	CG4603
603			FBgn0051279	CG31279
604			FBgn0014189	Hel25E
605			FBgn0052448	CG32448
606			FBgn0021764	sdk
607			FBgn0262843	CG43207

608			FBgn0010105	comm
609			FBgn0050269	CG30269
610			FBgn0037131	CG14564
611			FBgn0037552	CG7800
612			FBgn0029740	CG12680
613			FBgn0040849	Ir41a
614			FBgn0262877	CG43232
615			FBgn0085468	ND-MWFE
616			FBgn0033337	CG8272
617			FBgn0051704	CG31704
618			FBgn0051858	t-cup
619			FBgn0259227	CG42327
620			FBgn0050172	CG30172
621			FBgn0034214	CG6550
622			FBgn0050401	CG30401
623			FBgn0028379	fan
624			FBgn0036391	CG17364
625			FBgn0036494	Toll-6
626			FBgn0036360	CG10713
627			FBgn0032055	CG13091
628			FBgn0052750	CG32750
629			FBgn0003310	S
630			FBgn0027550	CG6495
631			FBgn0085489	CG34460
632			FBgn0037671	VhaM8.9
633			FBgn0051609	CG31609
634			FBgn0040651	CG15458
635			FBgn0085225	CG34196

636			FBgn0053688	CG33688
637			FBgn0038643	CG14300
638			FBgn0262005	CG42823
639			FBgn0085371	CG34342
640		28	FBgn0260011	NimC4
641			FBgn0083972	CG34136

Table A.1: Prediction result of type I transmembrane proteins: "Rank" means the rank of score using motif model, "APPL" means rank of homologous protein of Appl, same for "Notch".

Bibliography

[1] Annakaisa Haapasalo and Dora M Kovacs. The many substrates of presenilin/γ-secretase. *Journal of Alzheimer's Disease*, 25(1):3–28, 2011.

[2] John Hardy and Dennis J Selkoe. The amyloid hypothesis of alzheimer's disease: progress and problems on the road to therapeutics. *science*, 297(5580):353–356, 2002.

[3] Dennis J Selkoe and John Hardy. The amyloid hypothesis of alzheimer's disease at 25 years. *EMBO Molecular Medicine*, page e201606210, 2016.

[4] Jie Shen and Raymond J Kelleher. The presenilin hypothesis of alzheimer's disease: evidence for a loss-of-function pathogenic mechanism. *Proceedings of the National Academy of Sciences*, 104(2):403–409, 2007.

[5] Carlos A Saura, Se-Young Choi, Vassilios Beglopoulos, Seema Malkani, Dawei Zhang, BS Shankaranarayana Rao, Sumantra Chattarji, Raymond J Kelleher, Eric R Kandel, Karen Duff, et al. Loss of presenilin function causes impairments of memory and synaptic plasticity followed by age-dependent neurodegeneration. *Neuron*, 42(1):23–36, 2004.

[6] Georg Dietzl, Doris Chen, Frank Schnorrer, Kuan-Chung Su, Yulia Barinova, Michaela Fellner, Beate Gasser, Kaolin Kinsey, Silvia Oppel, Susanne Scheiblauer, et al. A genome-wide transgenic rnai library for conditional gene inactivation in drosophila. *Nature*, 448(7150):151–156, 2007.

[7] Gregory J Hannon. Rna interference. *Nature*, 418(6894):244–251, 2002.

[8] Satyajit Saurabh, Ambarish S Vidyarthi, and Dinesh Prasad. Rna interference: concept to reality in crop improvement. *Planta*, 239(3):543–564, 2014.

[9] Andrew Fire, SiQun Xu, Mary K Montgomery, Steven A Kostas, Samuel E Driver, and Craig C Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *nature*, 391(6669):806–811, 1998.

[10] Kishwar Hayat Khan. Gene expression in mammalian cells and its applications. *Adv Pharm Bull*, 3(2):257–63, 2013.

[11] Ashley J Pratt and Ian J MacRae. The rna-induced silencing complex: a versatile gene-silencing machine. *Journal of Biological Chemistry*, 284(27):17897–17901, 2009.

[12] Daniel Gautheret and André Lambert. Direct rna motif definition and identification from multiple sequence alignments using secondary structure profiles. *Journal of molecular biology*, 313(5):1003–1011, 2001.

[13] Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, 2003.

[14] Manolis Kellis, Nick Patterson, Bruce Birren, Bonnie Berger, and Eric S Lander. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *Journal of Computational Biology*, 11(2-3):319–355, 2004.

[15] Jaume Bacardit, Michael Stout, Jonathan D Hirst, Alfonso Valencia, Robert E Smith, and Natalio Krasnogor. Automated alphabet reduction for protein datasets. *BMC bioinformatics*, 10(1):1, 2009.

[16] Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

[17] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite: tools for motif discovery and searching. *Nucleic acids research*, page gkp335, 2009.

[18] Timothy L Bailey. Dreme: motif discovery in transcription factor chip-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

[19] Erik LL Sonnhammer, Gunnar Von Heijne, Anders Krogh, et al. A hidden markov model for predicting transmembrane helices in protein sequences. In *Ismb*, volume 6, pages 175–182, 1998.

[20] Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365, 1996.

[21] Johannes Söding, Andreas Biegert, and Andrei N Lupas. The hhpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, 33(suppl 2):W244–W248, 2005.

[22] Lusheng Wang and Tao Jiang. On the complexity of multiple sequence alignment. *Journal of computational biology*, 1(4):337–348, 1994.

[23] P Hogeweg and B Hesper. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *Journal of molecular evolution*, 20(2):175–186, 1984.

[24] David W Mount and David W Mount. *Bioinformatics: sequence and genome analysis*, volume 2. Cold spring harbor laboratory press New York:, 2001.

[25] Peter W Collingridge and Steven Kelly. Mergealign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC bioinformatics*, 13(1):117, 2012.

[26] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[27] Joseph Felsenstein, Stanley Sawyer, and Rochelle Kochin. An efficient method for matching nucleic acid sequences. *Nucleic Acids Research*, 10(1):133–139, 1982.

[28] Sanguthevar Rajasekaran, X Jin, and John L Spouge. The efficient computation of position-specific match scores with the fast fourier transform. *Journal of Computational Biology*, 9(1):23–33, 2002.

[29] Akira R Kinjo, Katsuhisa Horimoto, and Ken Nishikawa. Predicting absolute contact numbers of native protein structure from amino acid sequence. *Proteins: Structure, Function, and Bioinformatics*, 58(1):158–165, 2005.

[30] Burkhard Rost and Chris Sander. Conservation and prediction of solvent accessibility in protein families. *Proteins: Structure, Function, and Bioinformatics*, 20(3):216–226, 1994.

[31] Carl Ivar Branden et al. *Introduction to protein structure*. Garland Science, 1999.

[32] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. Weblogo: a sequence logo generator. *Genome research*, 14(6):1188–1190, 2004.

[33] Xiao-chen Bai, Eeson Rajendra, Guanghui Yang, Yigong Shi, and Sjors HW Scheres. Sampling the conformational space of the catalytic subunit of human γ-secretase. *eLife*, 4:e11182, 2016.

[34] Yanhui Hu, Ian Flockhart, Arunachalam Vinayagam, Clemens Bergwitz, Bonnie Berger, Norbert Perrimon, and Stephanie E Mohr. An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC bioinformatics*, 12(1):1, 2011.